Multi-Scale Long Short-Term Memory Network with Multi-Lag Structure for Blood Glucose Prediction

Tao Yang and **Ruikun Wu** and **Rui Tao** and **Shuang Wen** and **Ning Ma** and **Yuhang Zhao** and **Xia Yu** and **Hongru Li**[1]

Abstract. Accurate blood glucose (BG) prediction is necessary for daily glucose management of diabetes therapy. As glucose dynamics are often affected by various factors, such as diet, physical exercise, and insulin injection, it is difficult to consider all the relevant information and make a balance between the high-dimensional inputs and learning efficiency for a deep learning network. In this work, a novel multivariate predictor with a multi-scale long short-term memory (MS-LSTM) network was developed to automatically characterize the high-dimensional temporal dynamics and extract the features of blood glucose fluctuation and temporal trends sufficiently. Meanwhile, a multi-lag structure is designed for multiple variables, which can extract the dependence between different variables and blood glucose fluctuations more effectively. Furthermore, long-term sparse information was encoded and compressed to improve the learning efficiency of this deep learning network. The predictive capability of the proposed method was illustrated through 30-min and 60-min ahead glucose prediction in the OhioT1DM-2 Dataset. The root means square error (RMSE) values of 30-min and 60-min ahead predictions were 19.048 and 32.029, respectively, and the mean absolute error (MAE) values of 30-min and 60-min ahead predictions were 13.503 and 23.833. The results demonstrate the efficiency and prediction accuracy of the offline deep learning network, especially in the case of high-dimensional variables availability.

1 INTRODUCTION

Diabetes is a chronic disease characterized by the inability to maintain glucose homeostasis. Healthy pancreas controls the release of glucagon and insulin through α-cells and β-cells, respectively, to maintain normal blood glucose levels [7]. Type 1 diabetics cannot produce insulin normally because the β-cells are compromised, which leads to hyperglycemia and hypoglycemia [5], [17]. In recent years, advances in continuous glucose monitoring (CGM) and continuous subcutaneous insulin infusion (CSII) technologies have contributed to the closed-loop treatment of diabetes [1], [2], and [4]. The subcutaneous glucose concentration prediction algorithm has the potential to improve further the closed-loop treatment system for diabetes [8], [14], [15], and [18]. However, it is difficult to establish a multivariate physiological model to predict blood glucose precisely due to the influence of daily behaviors such as diet, physical exercise, and insulin injection [6]. Recently, some multivariate data-driven models are used to predict blood glucose levels and achieve satisfactory results. A successful case is the multivariable LSTM network proposed in paper [12], which has obtained better prediction results than the support vector regression model and diabetes experts.

[1] Northeastern University, China, email: lihongru@ise.neu.edu.cn

Nevertheless, different behaviors have different temporal effects on glucose fluctuation [3]. Using a unified lag for all variables may not be able to extract information about different characteristics sufficiently. Therefore, using multiple lags for each variable has positive implications for blood glucose prediction. An end-to-end recurrent neural network framework is proposed in paper [13], which is equipped with an adaptive input selection mechanism to improve the prediction performance of the multivariate time series. Based on this work, we develop a multi-scale LSTM (MS-LSTM) network that can capture the high-dimensional temporal dynamics and extract the features of blood glucose fluctuation and temporal trends sufficiently. Meanwhile, the multi-lag structure in the network can more effectively extract the dependence between different variables and blood glucose fluctuations. Compared with the traditional single-lag structure, using the multi-lag structure can extract more comprehensive features. Furthermore, long-term sparse information is encoded and compressed to accelerate the learning of deep networks. The MS-LSTM model was tested independently several times on the testing dataset, and the prediction results show that the model is excellent and robust.

This paper is organized as follows: section 2 explains the data preprocessing used; section 3 describes the architecture of the MS-LSTM network; section 4 illustrates model-free prediction methods in case of missing data; section 5 analyses the experimental results; section 6 summarizes the main contents from this study.

2 DATA PREPROCESSING

The variables selected for prediction included BG value, basal insulin dosage, bolus insulin dosage, carbohydrate intake, and timestamp [11]. Other variables provided were not selected for prediction, such as galvanic skin response, skin temperature, and acceleration. We used some data preprocessing methods, including aligning the original data, filling in the missing data, detecting and reconciling BG outliers, and normalizing the data. These data processing techniques will be illustrated in detail in the following sections.

2.1 Data alignment

The data in OhioT1DM-2 Dataset was collected by multiple devices, and some of the data was manually recorded by the patient, which caused the raw data to be asynchronous [16]. Therefore, the data needs to be aligned before feeding to the prediction model. Firstly, a time grid with a 5-minute sample period was derived based on the continuous glucose monitoring (CGM) data, and the missing data

was filled with zeros. Secondly, the timestamps of some insulin injections and carbohydrate intakes information cannot precisely match the timestamps of CGM data. They were reset to the timestamps of CGM data with the smallest time difference to keep the temporal correlation between the variables as much as possible [3].

2.2 CGM outlier detection and reconciliation

CGM measurements contain noise because of physical interference. Therefore, outlier detection and reconciliation are necessary to remove potential noise. Firstly, a gaussian process regression (GPR) model was trained to detect outliers of CGM measurements. The training dataset of the GPR model was the first 288 points of the training dataset. The input of the GPR model was CGM measurements from time $t-30$ to $t-5$, 6 points in total, and its output was mean ($\mu(t)$) and variance ($\sigma^2(t)$) of the CGM prediction at the time t. Then $\mu(t)$ and $\sigma^2(t)$ was used to reconcile CGM outlier at the time t as equation(1).

$$g(t) = \begin{cases} \mu(t) - 4.5\sigma^2(t) & , g(t) < \mu(t) - 4.5\sigma^2(t) \\ \mu(t) + 4.5\sigma^2(t) & , g(t) > \mu(t) + 4.5\sigma^2(t) \\ g(t) & ,others \end{cases} \quad (1)$$

where $g(t)$ is the BG level at time t.

2.3 Missing data filling

In the OhioT1DM-2 Dataset, basal insulin dosage and CGM measurements have missing data in some situations. As the basal insulin dosage has daily periodicity, it can be filled by the previous day's data. Although many methods are applied for missing CGM value filling, the accumulative error will inevitably increase as the number of filling increasing. Therefore, to degrade the accumulative error caused by data filling, the first-order Taylor series extrapolation and historical averages were weighted and summed to fill in the missing CGM values as the number of continuous missing items was less than 12. The respective methods for the missing numbers greater than or equal to 12 will be explained in detail later. It should be noted that the missing CGM values in the training dataset will not be filled to avoid additional noise.

2.4 Data normalization

Data normalization can accelerate deep network training and improve the accuracy of the model to a certain extent. We used three methods to normalize the data, and the results show that the model with coefficient normalization had the best performance. Coefficient normalization refers to only scale the amplitude of data to maintain the distribution of the raw data as much as possible [10]. The scaling of different variables was shown in Table 1.

Table 1. Scaling of different variables.

Variable	Glucose level	Timestamp	Basal	Bolus	Meal
Scaling	1/100	1/100	1/12	1	1/10

3 MS-LSTM MODEL

In this section, we will introduce the architecture of the MS-LSTM model and explain how the model is trained and tested.

3.1 Model architecture

As shown in Figure 1, the MS-LSTM model has a multi-scale hierarchy structure, which can learn the short-term and long-term dependence of blood glucose sequence. Meanwhile, the multi-lag structure can extract features on time-windows of different sizes, the features extracted on a large time-window are more abundant, and the features extracted on a small time-window are more time-sensitive. Therefore, compared with single-lag, the multi-lag structure can extract more comprehensive features and more effectively extract the dependence between different variables and blood glucose fluctuations. Theoretically, the more lags used, the more comprehensive features extracted, but correspondingly, the training time of the model will increase. Therefore, three lags were used for all variables to balance the training time and adaptability, as shown in Table 2, where PH represents the prediction horizon.

Table 2. Scale levels or lags of different variables.

	PH=30	PH=60
Variable	Scale level or lag	Scale level or lag
Blood glucose	1×7,2×7,3×7	1×9,2×9,3×9
Basal and timestamp	8,16,24	12,24,36
Bolus and timestamp	8,16,24	12,24,36
Meal and timestamp	8,16,24	10,20,30

Specifically, for predicting blood glucose after 30 minutes, the three scales adopted for the blood glucose variable were 1×7, 2×7, and 3×7, which means that all scale levels are 7, and the dilated sampling rate is 1, 2 and 3, respectively. Three lags of the basal variable were 8, 16, and 24, respectively. To ensure the unity of the output dimensions, in the multi-scale hierarchical and multi-lag structure, the number of LSTM states was equal to the minimum scale of blood glucose variable. As shown in Table 3, to sufficiently extract the useful information of various variables, the number of LSTM states in the feature fusion layer was 256. The number of nodes in the fully connected layer later was 256, 64, and 1, respectively, and some dropout layers are added between the fully connected layers to avoid the network overfitting problem.

Table 3. Detailed information of the MS-LSTM network.

		PH=30	PH=60
Structure	Layer name	Parameter	Parameter
Multi-scale hierarchical	LSTM	7 Unit	9 Unit
	LSTM	7 Unit	9 Unit
Multi-lag	LSTM	7 Unit	9 Unit
	LSTM	7 Unit	9 Unit
LSTM and Fully Connected (FC) layer	LSTM	256 Unit	256 Unit
	FC	256 Unit	256 Unit
	Dropout	0.2	0.2
	FC	64 Unit	64 Unit
	Dropout	0.1	0.1
	FC	1 Unit	1 Unit

3.2 Training and testing

The training data was divided into a training set and a verification set at a ratio of 9:1. The last 10% of the training dataset is closest to the

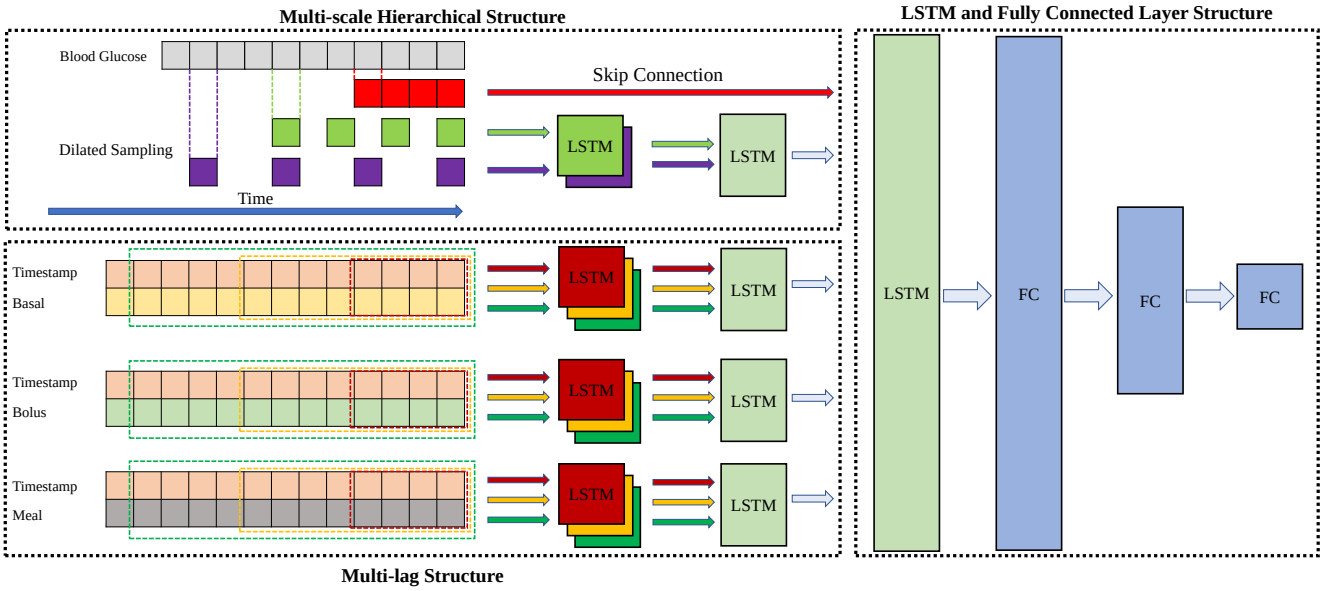

Figure 1. Block diagram representation of the MS-LSTM network

testing dataset in time, and its distribution is most similar to the testing dataset, so it was set apart as the verification set. When training the model, each iteration was evaluated on the verification set. When the model had not obtained better results after 300 consecutive evaluations, the training would be stopped, and the model which performs best on the verification set before would be saved. The training stop strategy that can effectively avoid the problem of overfitting the network is called early stopping. Because the 13th point on the test set needs to be predicted, some training data was added at the beginning of the test set to ensure that the number of prediction points meets the requirements. Besides, for several CGM data after a noticeable amount of continuously missing data, the model was not used for prediction. Instead, two model-free prediction algorithms with adaptive weight prediction and remain prediction were used to predict, respectively. Finally, the predictions were limited in the range of 40 to 400. The flow diagram is shown in Figure 2.

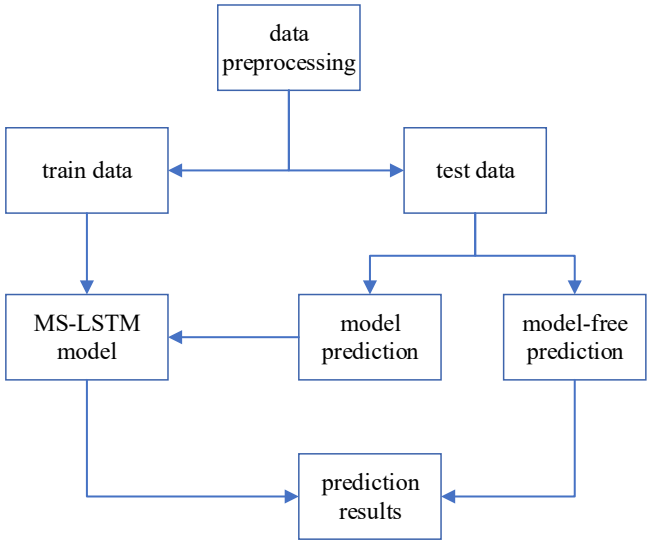

Figure 2. Flow diagram of blood glucose prediction

Training batch size: The experiment used mini-batch for weight adjustment, and the batch size of each update weight will affect the accuracy of the model. In the experiment, it was found that the larger batch size could improve the accuracy and accelerate the training process of the model, so the batch size was set to 1024.

Loss function: The experiment compared the negative log-likelihood (NLL) loss function, ε-insensitive loss function, mean absolute error (MAE) loss function, and root mean square error (RMSE) loss function. The results displayed that the model trained with the RMSE loss function had the best performance.

Optimizer: This experiment tested the root mean square prop (RMSProp) optimizer and adaptive moment estimation (Adam) optimizer [9]. The results showed that the performances of RMSProp and Adam were similar, but Adam had a significant advantage in the convergence speed. Therefore, Adam optimizer was used to update model weights, and the learning rate was set to 0.0001. In summary, the hyperparameters are shown in Table 4.

Table 4. Summary of the hyperparameters.

Hyperparameter	Value
Training batch size	1024
Optimizer	Adam optimizer
Learning rate	0.0001
Training stop strategy	early stopping
Loss function	RMSE

The experimental environment is Win10 Professional 64-bit operating system, the hardware platform is Intel Core i7 9750H processor, NVIDIA GeForce GTX 1660 Ti graphics processing unit, 16G memory notebook computer, and the development tool is Python 3.6, Keras 2.2.4, TensorFlow-GPU 1.12.0. The code used in the experiment is available on Github. In this hardware and software environment, the average training time for the MS-LSTM model was about 10 minutes.

4 MODEL-FREE PREDICTION

When the number of the missing CGM data is more than 11, the predictions of the MS-LSTM model for the following several values will cause a significant deviation. Therefore, for these CGM data, adaptive weight prediction and remain prediction are used instead of the model. The adaptive weight prediction algorithm uses short-term maintainability and long-term periodicity of blood glucose levels to make predictions. Specifically, when fewer CGM data are missing, the prediction is close to the last CGM value before the missing data, that is, depending on the short-term maintainability of blood glucose levels. On the contrary, when there are more missing data, the prediction is close to the CGM value at the same time of the previous day, that is, depending on the long-term periodicity of blood glucose levels. The process of adaptive weight prediction can be described by equation (2)-(4).

$$f = n_{miss}/(n_{miss}+c) \tag{2}$$

$$g_{av}(t) = \frac{1}{2n+1}\sum_{288-n}^{288+n} g(t - T \times n_{back}) \tag{3}$$

$$P_{aw} = (1-f)g_{last}(t) + f \times g_{av}(t)) \tag{4}$$

where n_{miss} is the number of missing data between the current prediction and the last CGM measurement before the missing data. c is a constant not less than 0, and the value in this experiment is set to 68. f is the adaptive weight factor, depend on n_{miss} and c. T is the blood glucose measurement period, the value in the OhioT1DM-2 Dataset is 5 minutes. n is a positive integer constant not less than 0, and the value in this experiment is set to 1. $n_{back} \in \{288-n, 288-n+1, ..., 288+n\}$. $g(t)$ is the BG level at time t. $g_{av}(t)$ is the average value of the CGM data of $2n+1$ points at the same time on the previous day, which represents the long-term periodicity of BG levels. $g_{last}(t)$ is the last CGM value before the missing data, which represents the short-term maintainability of BG levels. Finally, P_{aw} is the adaptive weight prediction value. As shown in Figure 3, the black points in the period from time D to F are the predictions produced by adaptive weight prediction algorithm.

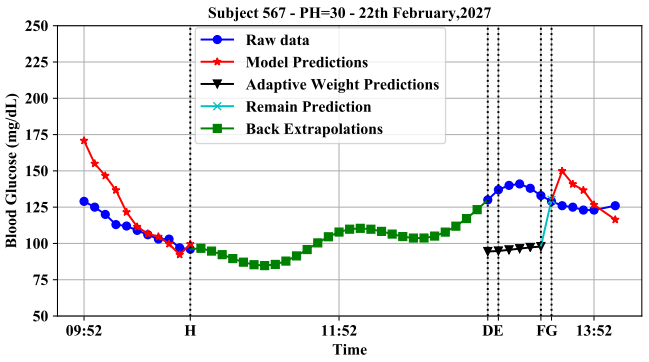

Figure 3. Prediction results of the three algorithms

When the first CGM data appears after the missing data, the value would be directly used as the predicted value of the required prediction horizon. So this algorithm is called remain prediction. As shown in the sky blue point in Figure 3, the blood glucose value at time D was the prediction value at time G.

Then, when two CGM values appeared after the missing data, as shown about the BG values at time D and E in Figure 3. Based on these two points, the reverse first-order Taylor series extrapolation was performed. Then the extrapolated data and the average historical data before the missing data were weighted and summed to ensure the smoothness of the filled data. The green points in Figure 3 were the extrapolated backward data, which were used by the MS-LSTM model to predict BG level after time G.

5 RESULTS AND ANALYSIS

The performance of the model was evaluated by the root mean square error (RMSE) and mean absolute error (MAE) between the predictions and the original test data.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{6}$$

where $\hat{y}_i$ is the predicted BG value, y_i is the target value and N represents the size of the testing dataset. To be noted that, the extrapolated values of BG were removed when evaluating the performance of the model, which guarantees the predictions had the same number as the test data.

According to the preceding steps, the results of four independent experiments are summarized in Table 5, where SD represents the standard deviation. All subjects used the same experimental parameters, but the RMSE of each patient varied from 15 to 22. Among them, the smallest RMSE is 15.871 for patient 596, and the largest RMSE is 21.934 for patient 567. The prediction results are shown in Figure 4-5. It is worth noting that the average RMSE variance of the MS-LSTM model is only 0.061 in 30 minutes prediction horizon, which reflects the excellent robustness of the model.

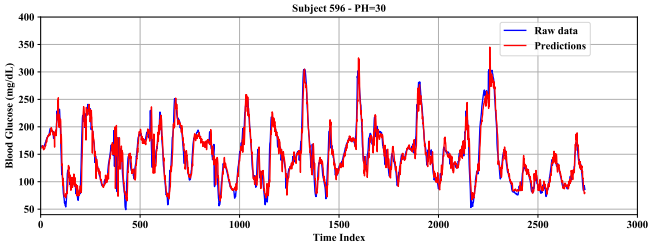

Figure 4. Blood glucose prediction results of subject 596 produced by the MS-LSTM model

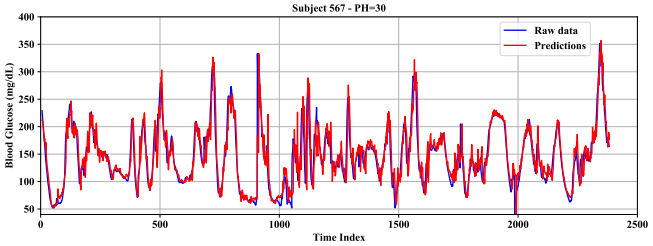

Figure 5. Blood glucose prediction results of subject 567 produced by the MS-LSTM model

The subject 567 has many consecutive spikes, which is the primary source for the prediction error. Besides, another source of prediction

Table 5. RMSE and MAE values of the MS-LSTM model for 6 subjects.

Subject	Test point	PH=30		PH=60	
		Average RMSE ± SD	Average MAE ± SD	Average RMSE ± SD	Average MAE ± SD
540	2884	20.996 ± 0.062	15.244 ± 0.051	38.219 ± 0.029	28.675 ± 0.017
544	2704	16.687 ± 0.025	11.679 ± 0.014	27.424 ± 0.100	19.522 ± 0.030
552	2352	16.918 ± 0.064	12.726 ± 0.058	30.109 ± 0.185	23.340 ± 0.320
567	2377	21.934 ± 0.039	14.698 ± 0.027	37.155 ± 0.369	27.324 ± 0.377
584	2653	21.881 ± 0.142	15.417 ± 0.127	33.913 ± 0.026	25.362 ± 0.091
596	2731	15.871 ± 0.035	11.258 ± 0.041	25.358 ± 0.227	18.777 ± 0.137
Mean		19.048 ± 0.061	13.503 ± 0.053	32.029 ± 0.156	23.833 ± 0.162

error is the missing data, as shown in the predictions after the missing data in Figure 6. Finally, a slight time delay is observed in the prediction curve, and it is also a problem for most prediction methods.

The CGM measurements contain noise because of physical interference. We used the GPR model to detect and reconcile CGM outliers to the greatest extent. However, only some severe outliers were detected and reconciled because there was no judgment standard for outliers. There are still many outliers in the raw CGM data, which is very unfavorable for the prediction model learning. Therefore, denoising CGM and obtaining high-quality data is very important to improve the performance of the prediction model.

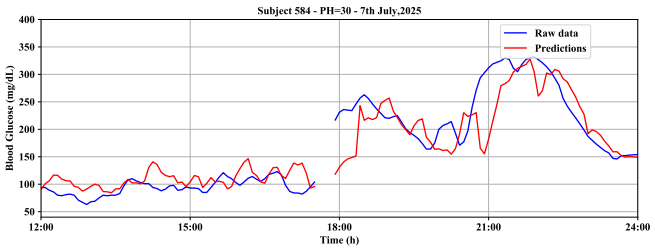

Figure 6. Prediction performance in case of missing data

6 CONCLUSION

In this paper, the MS-LSTM network is developed to adaptively characterize high-dimensional temporal dynamics and extract the long-term and short-term features of glucose fluctuation. Meanwhile, a multi-lag structure is designed for multiple variables, which can extract the dependence between different variables and blood glucose fluctuations more effectively. The long-term sparse temporal data is encoded and compressed to suitable for efficient learning with the model. The mean value of the RMSE for 6 subjects is 19.048, with standard deviation equals to 0.061 in 30-minute PH. Missing data and rapid fluctuations in blood glucose levels are the two main factors that affect the prediction performances of the model.

7 FUNDING

This research was supported by National Natural Science Foundation of China (No.61973067 and No.61903071).

REFERENCES

[1] R.M. Bergenstal, D.C. Klonoff, S.K. Garg, B.W. Bode, M. Meredith, R.H. Slover, A.J. Ahmann, S.W. Welsh, J.B. Lee, F.R. Kaufman, and AI-HS Group, 'Threshold-based insulin-pump interruption for reduction of hypoglycemia', *New England Journal of Medicine*, **369**, 224–232, (2013).

[2] B. Buckingham, F. Cameron, P. Calhoun, D.M. Maahs, D.M Wilson, H.P. Chase, B.W. Bequette, J. Lum, J. Sibayan, and R.W. Beck, 'Outpatient safety assessment of an in-home predictive low-glucose suspend system with type 1 diabetes subjects at elevated risk of nocturnal hypoglycemia', *Diabetes Technology & Therapeutics*, **15**, 622–627, (2013).

[3] J.W. Chen, K.Z. Li, P. Herrero, T.Y. Zhu, and P. Georgiou, 'Dilated recurrent neural network for short-time prediction of glucose concentration', in *CEUR Workshop Proceedings*, volume 2148, pp. 69–73, Stockholm, Sweden, (2018).

[4] P. Dua, F.J. Doyle, and E.N. Pistikopoulos, 'Multi-objective blood glucose control for type 1 diabetes', *Medical & Biological Engineering & Computing*, **47**, 343–352, (2009).

[5] A. Facchinetti, S. Favero, G. Sparacino, and C. Cobelli, 'An online failure detection method of the glucose sensor-insulin pump system: Improved overnight safety of type-1 diabetic subjects', *IEEE Transactions on Biomedical Engineering*, **60**, 406–416, (2013).

[6] E.I. Georga, V.C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni, D. Polyzos, and D.I. Fotiadis, 'Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression', *IEEE Journal of Biomedical and Health Informatics*, **17**, 71–81, (2013).

[7] N.D.D. Group, 'Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance', *Diabetes*, **28**, 1039–1057, (1979).

[8] C.S. Hughes, S.D. Patek, M.D. Breton, and B.P. Kovatchev, 'Hypoglycemia prevention via pump attenuation and red–yellow–green "traffic" lights using continuous glucose monitoring and insulin pump data', *Journal of Diabetes Science & Technology*, **4**, 1146–1155, (2010).

[9] D.P. Kingma and J.L. Ba, 'Adam: A method for stochastic optimization', in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, San Diego, CA, United states, (2015).

[10] J. Martinsson, A. Schliep, B. Eliasson, C. Meijner, S. Persson, and O. Mogren, 'Automatic blood glucose prediction with confidence using recurrent neural networks', in *CEUR Workshop Proceedings*, volume 2148, pp. 64–68, Stockholm, Sweden, (2018).

[11] C. Midroni, P. J. Leimbigler, G. Baruah, M. Kolla, A.J. Whitehead, and Y. Fossat, 'Predicting glycemia in type 1 diabetes patients: Experiments with xgboost', in *CEUR Workshop Proceedings*, volume 2148, pp. 79–84, Stockholm, Sweden, (2018).

[12] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz, 'Using lstms to learn physiological models of blood glucose behavior', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2887–2891, Jeju Island, Korea, Republic of, (2017).

[13] L. Munkhdalai, T. Munkhdalai, H.P. Kwang, T. Amarbayasgalan, E. Erdenebaatar, Hyun W.P., and Keun H.R., 'An end-to-end adaptive input selection with dynamic weights for forecasting multivariate time series', *IEEE Access*, **7**, 99099–99114, (2019).

[14] M. Phillip, T. Battelino, E. Atlas, O. Kordonouri, N. Bratina, S. Miller, T. Biester, S.M. Avbelj, I. Muller, R. Nimri, and T. Danne, 'Nocturnal glucose control with an artificial pancreas at a diabetes camp', *New England Journal of Medicine*, **368**, 824–833, (2013).

[15] J.C. Pickup, 'Insulin-pump therapy for type 1 diabetes mellitus', *New England Journal of Medicine*, **366**, 1616–1624, (2012).

[16] J.Y. Xie and Q. Wang, 'Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge', in *CEUR Workshop Proceedings*, volume 2148, pp. 97–102, Stockholm, Sweden, (2018).

[17] S. Zavitsanou, A. Mantalaris, M.C. Georgiadis, and E.N. Pistikopoulos, 'In silico closed-loop control validation studies for optimal insulin delivery in type 1 diabetes', *IEEE Transactions on Biomedical Engineering*, **62**, 2369–2378, (2015).

[18] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, 'Reduction of number and duration of hypoglycemic events by glucose prediction methods: a proof-of-concept in silico study', *Diabetes Technology & Therapeutics*, **15**, 66–77, (2013).